Introduction - Heart Disease in Cleveland

According to the CDC, one person dies from heart disease every 34 seconds in the United States alone with about 1 in 7 deaths attributed to it. As such, any sort of advantage that can be gained must be had to fight this disease. Our analysis helps identify attributes that are key in predicting if someone has heart disease or not. Doctors would be able to take these data supported considerations when identifying patients who are at the moment undiagnosed. Our data represent over 300 individuals from the Cleveland area in 1988 of which it is known if they have heart disease or not. We hope that our analysis would be used to further preventative measures taken by doctors and patients to curb this deadly disease.

Presentation and GitHub

Methods

In order to properly use the data, we had to aggregate all four types of heart disease into a dichotomous scale of heart disease being present or not. As part of the process of finding the nearest neighbors, the data were also standardized. We used the nearest neighbors to determine the possibility of a novel individual having heart disease. This is appropriate as there are often cohorts amongst society that make heart disease more prevalent in that group.

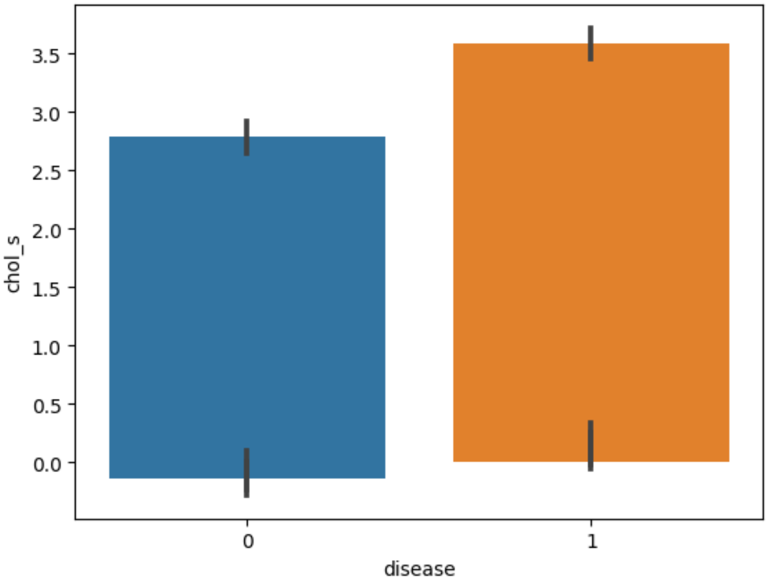
The general process of finding a near neighbor is computing the distance between the initial observation and other observations. We chose the l-2 or Euclidean measure of distance, which would give us the distance between the points in higher dimensional space; where the dimension is how many attributes we selected.

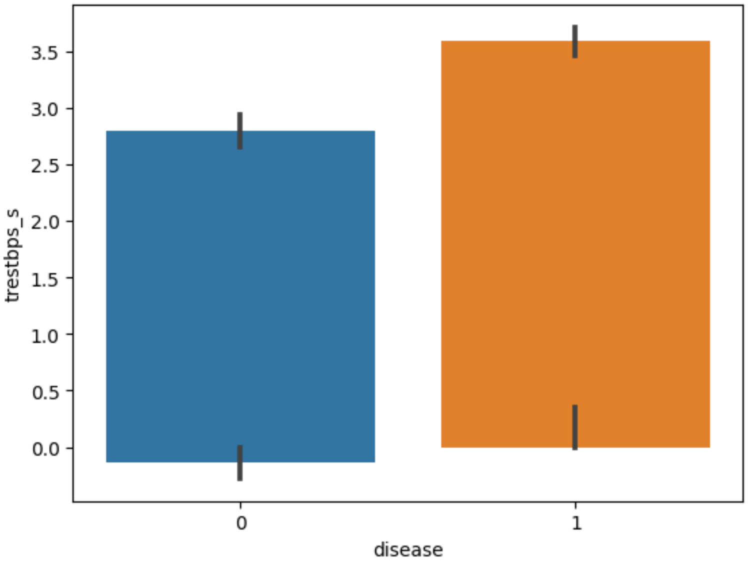
To verify that our model and assumptions were valid, the first thing that we did was to report on how the model did with a random individual. This we did by verifying that the attributes selected were indeed predictive of the presence of heart disease. Naturally, just one individual result on a model isn't enough, so we also increased the number of patients modeled, and then cross validated the model.

Part of the K-nearest neighbors is to cross validate the model, or to ensure its robustness and generalizability by segregating and executing the model on the test data. For example, a set of training data is partitioned randomly into tenths for Monte Carlo Validation, where the model is trained on 90% of the data then tested on the remaining tenth. This ensures the models generalizability by assuring us that the strength of the model is not due to random chance.

Results

In our analysis of heart disease, we found that there were just a few attributes that predicted the presence of heart disease in a novel observation. What we found mostly corroborated the general perception of heart disease. That is, things like age, resting heart rate, chest pain and severity or existence of thalassemia (a change in blood flow in the heart), were great at predicting the presence of heart disease. Another attribute that helped predict the disease was if exercise induced a type of heart pain in the patient. We found that by using these five attributes, we were able to correctly predict 80% of the cases (recall) and have just 14% of false negatives (precision).





We also tested how the number of neighbors used in our model affected our ability to predict. We found that often the more neighbors used in predicting, the better the score we achieved but that improvement curved pretty quickly. After 7 or so neighbors the improvement is unnoticeable. Along with neighbor count we found that the number of attributes chosen affected the score. Having too few attributes gave inconsistent results when testing with monte carlo cross validation. Having too many resulted in confounding variables affecting how the model predicted lowering average score. So those 5 attributes we found the most useful.

Introduction - Power Consumption in Morocco

The power market is a 1.4 trillion dollar industry. One major complexity is that once electricity is generated at the plant, it must be used. This leaves the electric company vulnerable to over supplying energy and wasting valuable resources. Especially during power surges (such as when people get home from work, during a hot day, or first thing in the morning), the grid needs to respond to every change, with the degree of accuracy connected to the savings of the company. As such, millions of dollars are invested into predicting when these power spikes happen. We seek to describe when the above average usage of power will happen based on similar time periods in the past; using the nearest neighbors of a time slot to predict the future.

Dataset

The data on power consumption in Tetouan was obtained from the UCI Machine Learning Repository and is from 2017. The data describe the various weather conditions (temperature, humidity, wind speed, and diffuse flows) and the power consumption for each zone of the city. For demonstration purposes, we used all available attributes and predicted high power usage in zone 1. Data was obtained every 10 minutes for a year.

Methods

The methods used in analyzing these data are identical to those used in the analysis of heart disease. The only consideration to be made would be that we also had to classify if a particular time of power usage was above average or not, which we did by finding the deviation from the mean power usage for that zone.

Results

In our power consumption for the city of Tetouan, Morocco, we found that by utilizing nearly all of the available attributes, we were able to predict quite well the energy usage for Zone 1. These attributes are humidity, diffuse air flows and the nearby power consumption. These are important because they can be measured, and then our model can predict in real time if more energy is needed. We found that by using these attributes, we were able to correctly predict 82% of the above average power usages (recall) and have just 8% of false negatives (precision).

Since the data had plenty of observations, we were able to increase the number of neighbors without a loss of generalizability. Accordingly, we looked at the nearest 100 neighbors to get our results. All in all, we were able to learn and predict quite a lot from the nearest times across the year for Tetouan.